Weekly Report 06/01/2014

The data inspection project

Quartile and quantile functions implemented at the server end.
Not much progress since Wanqi Hu is working on his thesis. I'm still waiting for his return.

VASA Project

No progress since Jieqiong is refactoring the project codes.

Routing Evaluation Project

This is the project I'm working with Jieqiong for her CHI paper. We evaluate all visual enhancements for routing path recognition with proper user study. We've finalized the tasks and working on IRB proposals. Implementation is in progress.

VAST Project

We're working on VAST challenges, Jiawei on the third one and me on the first and the second one (The proposal is attached at the end). We've imported the dataset into Jigsaw for initial data inspection. After that, we'll come up with a visualization design to present our findings.
Next Monday we will have a meeting with Victor Chen's group about the collaboration of the second challenge.

Tool Learning

I installed and learned the ArcGIS[1] software, the powerful visualization and analytics tool of geo-located data. It's not free but Purdue University has their license key. They also have APIs of Javascript, Java, Python, C# …
GIS service is useful in the VASA project, VAST project as well as the data cleaning project.

Miscellaneous

1. Paper review (VACCINE Lab)
   NUMARCK: Machine Learning Algorithm for Resiliency and Checkpointing
   This paper introduced a machine learning algorithm for light-weighted data resiliency and checkpointing in the HPC environment. They adopted the strategy in video processing, capturing only data of change, to minimize data storage in checkpointing. They also put forward a new data approximation method, which learns the number of data clusters with distribution-based binning.

[1] http://www.esri.com/software/arcgis

APPENDIX

Mini Challenge 1

Mini-Challenge 1 focuses on the disappearance. As an analyst, your task is to bring law enforcement up to date on the current organization of the POK and how that organization has changed over time, as well as to characterize the events surrounding the disappearance.

Questions for Participants

- Provide a visual representation of the structure of the Protectors of Kronos network, with supporting evidence.
    - Who are the leaders?
    - Who is part of the extended network?
    - How has the group structure and organization changed over time?
    - Where are the potential connections between the POK and GAStech?
- Describe the events of January 20-21, 2014. What is the timeline of events?
- Identify at least two possible explanations why the GAStech employees may be missing. What evidence do you have to support each of these explanations?

Available Data

- A map of Kronos
- A chart describing the local GAStech organization.
- A spreadsheet of GAStech employee records. The primary worksheet contains the data; the index worksheet contains the data dictionary.
- Email headers from two weeks of internal GAStech company email, in comma-separated values (CSV) format
- Resumes and short biographies of many, but not all, of the GAStech employees, in Microsoft Word format
- Historical reports and descriptions of the countries involved, in Microsoft Word format
- Relevant current and historical news reports from multiple domestic and translated foreign sources, in text file format. Because these articles have come from multiple sources and original formats, some of them may contain corrupted characters, which is typical for this type of data. These corrupted characters should not interfere with your ability to analyze the data.

Our proposal

For the short news and long reports, we use Jigsaw to extracting keywords and timeline. For

the email communication data, we build up our own email network to analyze it. Combining the two results, we composite a story and design a poster of a complete story together with Dr. Victor Chen.

Approaches

- Use Jigsaw to analyze the short and long reports of Kronos, collecting keywords of persons, organizations, events and dates.
- Construct a network with the email dataset, applying time slider for filtering. Employee information will be the complementary information for each node in the network.
- Integrate the information from both sides, coming up with a solution/story, working with Dr. Victor Chen to design a poster for the story.

Mini Challenge 2

Questions for Participants

- **MC2.1** – Describe common daily routines for GAStech employees. What does a day in the life of a typical GAStech employee look like?
- **MC2.2** – Identify up to twelve unusual events or patterns that you see in the data. For each pattern or event you identify, describe
    - What is the pattern or event you observe?
    - Who is involved?
    - What locations are involved?
    - When does the pattern or event take place?
    - Why is this pattern or event significant?
    - What is your level of confidence about this pattern or event? Why?
- **MC2.3** – Like most datasets, the data you were provided is imperfect, with possible issues such as missing data, conflicting data, data of varying resolutions, outliers, or other kinds of confusing data. Considering MC2 data is primarily spatiotemporal, describe how you identified and addressed the uncertainties and conflicts inherent in this data to reach your conclusions in questions MC2.1 and MC2.2.

Available Data

- A list of vehicle assignments by employee, in CSV format (car-assignments.csv)
    - Employee Name
    - Car ID (integer)
    - Current Employment Type (Department; categorical)
    - Current Employment Title (job title; categorical)

- ESRI shapefiles of Abila and Kronos (in the Geospatial folder)
- A CSV file of vehicle tracking data (gps.csv) (a 14-day dataset)
    - Timestamp
    - Car ID (integer)
    - Latitude, Longitude
- A CSV file containing loyalty card transaction data (loyalty_data.csv) (a 14-day dataset)
    - Timestamp
    - Location (name of the business)
    - Price (real)
    - Name
- A CSV file containing credit and debit card transaction data (cc_data.csv) (a 14-day dataset)
    - Timestamp
    - Location (name of the business)
    - rice (real)
    - Name
- A tourist map of Abila with locations of interest identified, in JPEG format

Our proposal

Uncertainties and Conflicts definition --- For the second challenge, the uncertainties and conflicts would be mismatch of car GPS location and employees' transaction data. Given a 14-day trajectory and card transaction, we will compare everyone's daily pattern. So uncertainties and conflicts may also include mismatch of one's certain day trajectory with his/her daily routine pattern.